Neurocomputing 180 (2016) 35–54

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An object-based visual selection framework

Alcides X. Benicasa [a,*], Marcos G. Quiles [b], Thiago C. Silva [c], Liang Zhao [c], Roseli A.F. Romero [c]

[a] Department of Information Systems, Federal University of Sergipe (UFS), Itabaiana, Sergipe, Brazil
[b] Department of Science and Technology, Federal University of São Paulo (UNIFESP), São Paulo, SP, Brazil
[c] Department of Computer Science, Institute of Mathematics and Computer Science, University of São Paulo (USP), São Carlos, SP, Brazil

ARTICLE INFO

Article history:
Received 30 March 2015
Received in revised form
6 October 2015
Accepted 6 October 2015
Available online 27 November 2015

Keywords:
Bottom-up and top-down visual attention
Object-based attention
Location-based attention
Image segmentation
Recognition of objects
Low level and high-level classification

ABSTRACT

Real scenes are composed of multiple points possessing distinct characteristics. Selectively, only part of the scene undergoes scrutiny at a time, and the mechanism responsible for this task is named selective visual attention. Spatial location with the highest contrast might highlight from scene reaching level of awareness (bottom-up attention). On the other hand, attention may also be voluntarily directed to a particular object in the scene (object-based attention), which requires the recognition of a specific target (top-down modulation). In this paper, a new visual selection model is proposed, which combines both early visual features and object-based visual selection modulations. The possibility of the modulation regarding specific features enables the model to be applied to different domains. The proposed model integrates three main mechanisms. The first handles the segmentation of the scene allowing the identification of objects. In the second one, the average of saliency of each object is computed, which provides the modulation of the visual attention for one or more features. Finally, the third builds the object-saliency map, which highlights the salient objects in the scene. We show that top-down modulation has a stronger effect than bottom-up saliency when a memorized object is selected, and this evidence is clearer in the absence of any bottom-up clue. Experiments with synthetic and real images are conducted, and the obtained results demonstrate the effectiveness of the proposed approach for visual selection.

1. Introduction

Every day we face complex scenes, and our visual system needs to analyze and understand a large amount of visual information while ignoring unimportant things. To handle this information, our visual system must deliver proper attention to specific objects, this task is called visual attention or visual selection. Biologically speaking, an object with features that contrast to the background pops out and draws attention automatically [1,2]. The information that defines the contrast among objects is related to both primitive features of the scene and previous knowledge about specific targets (memory). This information is related to two distinct components that drive the human visual attention. The first one is the *bottom-up* attention, which is involuntarily guided by visual features (such as colors, depth, and motion). The second is the *top-down* modulation, which voluntarily guides the attention towards specific features or known objects in the scene [3]. It is worth noting that when there are objects with similar primitive features, top-down modulation turns out to be a dominant attribute to select one of the objects as a target [4].

But what is an object? An object can be defined as anything that is visible or tangible and is relatively stable in form. For simplicity, here each object is referred to a segment segregated from the ground of the scene. Moreover, according to Temporal Correlation theory, we might say that an object is a binding of distinct features gathered from the visual scene into a single percept. Behavioral and neurophysiological evidences have shown that the selection of objects is used in the primate visual system [5,6]. It is believed that a pre-attentive process or a perceptual organization is run by the brain unconsciously, performing a figure-ground segregation and segmentation of the visual scene in a set of objects. Those objects, in fact, compete for attention [7]. The Perceptual Organization has been studied in Gestalt psychology where it is indicated that the world is perceived as a cluster of well-structured objects and not as a collection of unorganized points. The formation of objects is governed by Gestaltian laws of grouping such as connectivity, proximity and similarity. It is worth noting that although these processes can be characterized as *bottom-up*, they can also be influenced by *top-down* mechanisms [4].

The *Temporal Correlation Theory* [8,9] offers an interesting approach to representing multiple objects in a scene by using artificial neural network models, such as the LEGION (Locally

* Corresponding author.
E-mail addresses: alcides@ufs.br (A.X. Benicasa), quiles@unifesp.br (M.G. Quiles), thiagoch@icmc.usp.br (T.C. Silva), zhao@usp.br (L. Zhao), rafrance@icmc.usp.br (R.A.F. Romero).

http://dx.doi.org/10.1016/j.neucom.2015.10.111
0925-2312/

Excitatory Globally Inhibitory Oscillator Network) [10,11]. Thus, by taking the temporal correlation into account, we can integrate distinct features and deal with several objects in a scene.

Bottom-up models [12–15], do not consider the role of a working memory in the visual selection. In those models, only the primitive features of the image are used to identify the salient point or region of the scene. The selection is directly related to the unsupervised learning, whose goal is to find groups of similar objects according to their features without supervision, or involuntarily. On the other hand, the associative memory considered in top-down models might be associated with some form of supervision [16]. In this case, the visual system searches for previously known objects, which is an inherent characteristic of supervised learning methods. This processing involves a concept of working memory, which temporally holds some information about a target used to modulate the selection process. It is worth noting that top-down attention might also influence the response of bottom-up clues. According to [17,18], bottom-up attention alerts us to salient details in the scene, whereas top-down attention modulates bottom-up signals to bias the features of a specific target.

In visual selection research that considers both bottom-up and top-down modulations, one can observe two different approaches: (1) the development of psychophysical/psychological or computational models to reproduce real vision mechanisms and/or effects; and (2) the development of pattern recognition tools that simulates biological visual selection mechanisms, i.e., how to use memory for pattern storage and recall. Many works have been developed in this category [19–26].

In the first approach, previous works developed so far have focused on describing the biological/psychological mechanisms of working memory in the process of visual selection [27–30]. However, they still lack the definition of computational models that properly address this issue. For this reason, we propose a computational model to study how working memory can influence or benefit visual selection. In this sense, our paper makes a contribution to the first approach of visual selection research.

Specifically, a new object-based visual selection model with both bottom-up and top-down modulations is proposed. Thus, our model is composed of the following modules: (1) Visual Feature Extraction module responsible for extracting the early visual features, such as, colors and orientation; (2) LEGION network [10], for the image segmentation; (3) Network-Based High Level Data Classification, named HLC [31], for object recognition; (4) Network of Integrate and Fire Objects, which creates our object-saliency maps, and finally; (5) Object Selection Module that selects all the salient objects in the scene, based on the guidance from the object-saliency map.

By using the LEGION network, we provide an elegant manner to code temporally the objects in the scene [11]. It means that the objects formed during the segmentation process are highlighted one at a time, which allows a serial scanning of the visual scene. Moreover, the LEGION network is a well-known model consistent with the temporal correlation theory, and it has been extensively analyzed and applied to several tasks [10,11]. Also, our model considers prior external information about an object by combining the low-level and high-level data classifications [31]. The high-level classification exploits the complex topological properties of the underlying network constructed from the input data. We show that the combined classification approach is robust against changes in pattern recognition.

By integrating the modules mentioned above and using our new object-saliency map our model can delivery attention to objects of the scene regarding their visual features or previous knowledge of the objects/domain. Additionally, we provide qualitative and quantitative comparisons of the proposed model against ground truth fixation maps [32] and nine state-of-the-art methods employed for salient detection [12,33–40] for salient detection.

In summary, the main contributions of this work are:

1. The salience value of an object is considered one of the major components of this work. It is calculated using bottom-up features, top-down features, or both.
2. The absence of specific salient features may automatically ignore regions or objects in the scene.
3. Objects with a value of saliency below of the threshold value will not participate in the competition for attention;
4. The object-salience map is defined as a network composed of objects, with two types of connections, excitatory and inhibitory, responsible for synchronizing groups of objects that represent close patterns of similarities, and to inhibit objects related to background objects of the scene, respectively, allowing the object related to the most salient object of the scene to be selected.

This paper is organized as follows. In Section 2, a brief review of the early visual features extraction, the segmentation model, and the network-based high-level data classification is provided. Section 3 introduces the proposed model. Computer simulations are presented in Section 4. Finally, concluding remarks and future directions are drawn in Section 5.

2. Background

In this section, some feature combination strategies are reviewed, the segmentation mechanism and the network-based high-level data classification used in the visual selection model proposed.

2.1. Extraction of early visual features

The first step of the visual system consists in extracting the primitive information of the input image. According to [3], the first processing stage in any model of bottom-up attention is the computation of early visual features across the entire visual scene, where neurons at the earliest stages are tuned to simple visual attributes. In this work, feature maps are used to detect local spatial discontinuities in intensity contrast, color, orientation, size and location. Moreover, top-down modulations are also used to create maps for objects recognition.

According to [3], the saliency map is produced initially by a set of maps representing primary features, such as intensity, color, and orientation, extracted from the input scene. After that, to model the center-surrounding receptive fields, operations are performed on different spatial scales of those maps. This process, followed by a normalization operator, results in a new set of maps named feature maps. Next, these feature maps are combined into a set of conspicuity maps. Finally, the saliency map is obtained from a linear combination of the conspicuity maps. In this work, only the conspicuity maps are considered and used to build our object-saliency map.

Formally and according to [41], an input image $\mathbf{I}$ is sub-sampled into a dyadic Gaussian pyramid by convolution with a linearly separable Gaussian filter and decimation by a factor of two. This process is repeated to obtain the next levels $\sigma = [0, \ldots, 8]$ of the pyramid. According to [42], the encoding process is equivalent to sampling of the image with Laplacian operators of many scales. Thus, the code tends to enhance salient image features.

The intensity contrast is a spatial difference in light intensity in an image [3]. The intensity map I for each level (i) of the pyramid $\mathbf{I}$